Analytical and Decision Support Tools for Genomics-Assisted Breeding

Trends in Plant Science

CellPress

Review

Analytical and Decision Support Tools for Genomics-Assisted Breeding

Rajeev K. Varshney,[1,2,@,*] Vikas K. Singh,[1] John M. Hickey,[3] Xu Xun,[4] David F. Marshall,[5] Jun Wang,[4] David Edwards,[2] and Jean-Marcel Ribaut[6]

To successfully implement genomics-assisted breeding (GAB) in crop improvement programs, efficient and effective analytical and decision support tools (ADSTs) are 'must haves' to evaluate and select plants for developing next-generation crops. Here we review the applications and deployment of appropriate ADSTs for GAB, in the context of next-generation sequencing (NGS), an emerging source of massive genomic information. We discuss suitable software tools and pipelines for marker-based approaches (markers/haplotypes), including large-scale genotypic and phenotypic, data management, and molecular breeding approaches. Although phenotyping remains expensive and time consuming, prediction of allelic effects on phenotypes opens new doors to enhance genetic gain across crop cycles, building on reliable phenotyping approaches and good crop information systems, including pedigree information and target haplotypes.

Trends

Appropriate analytical and decision support tools (ADSTs) are critical for deploying genomics-assisted breeding.

Development of breeder-friendly pipelines and/or tools will enhance the adoption of ADSTs and facilitate the rapid development of new breeding lines.

Deployment of ADSTs in public breeding programs is the need of the hour.

Advances in next-generation sequencing technologies have prompted geneticists and breeders to utilize more sophisticated tools for sequencing-based mapping and genome-wide selection for the development of new breeding lines.

The availability of open-source and one-stop integrated platforms such as Integrated Breeding Platform (IBP) and their hubs across the world will facilitate the modernization of crop breeding programs.

Breeding for Sustainable Food Production

GAB (see Glossary) has become popular for crop improvement in recent years partly due to availability of low-cost **high-throughput genotyping (HTPG)** and NGS technologies. Several successful examples of GAB are now available not only in major crop species but also in many so-called 'orphan crops' [1,2]. GAB pipelines involve various steps including: characterization of diverse germplasm collections; development of mapping populations; identification of genomic regions through genetic or association mapping; and application of markers in breeding. Numerous **ADSTs** are required throughout all four of these steps [3]. Better understanding of the genetic diversity that is present in germplasm collections in gene banks and breeding material helps breeders identify new valuable alleles for breeding. Field evaluation of large germplasm collections is challenging due to, for example, poor genetic background, variation in phenology, the logistics and resources required, and selection of smaller subsets that represent the diversity of the collection. These sets include '**core collections**' (10% of the entire collection) [4], '**mini-core collections**' (about 10% of the core collection or 1% of the entire collection) [5], and '**reference sets**' (usually developed based on the molecular characterization of a composite collection) [6]. Efforts to define these sets should also benefit from the use of ADSTs on germplasm collections.

For trait mapping, two complementary approaches – namely, linkage mapping and association mapping, which in the context of large-scale genotyping and the **whole-genome re-sequencing** era are now referred to as **genome-wide association studies (GWAS)** – have been used in crop genetics. Construction of high-quality **genetic maps** with precise marker orders is critical when undertaking **quantitative trait locus (QTL)** analysis, which leads to the identification of genomic

[1]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India
[2]School of Plant Biology, University of Western Australia, 35 Stirling Highway, Crawley, Australia
[3]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, UK
[4]Beijing Genomics Institute (BGI) Shenzhen, Shenzhen, China
[5]Information and Computational Sciences, The James Hutton Institute, Invergowrie, Dundee, UK

[6]Generation Challenge Program/Integrated Breeding Platform, c/o CIMMYT, Apdo. Postal 6-641, DF Mexico, Mexico

*Correspondence: r.k.varshney@cgiar.org (R.K. Varshney).
[@]Twitter: @rajvarshney

regions and markers associated with target traits. Association mapping or GWAS has emerged as a new approach for the identification of causal loci/genes for traits of interest and some tools have been developed in recent years. Availability of the re-sequencing data of multiple accessions of the same species or different species has initiated the concept of the pan-genome. The generated hapmap information through pan-genome analysis is useful for the construction of high-density linkage maps. Pan-genomes are useful for the collection of all the genes at clad level. Additionally, with re-sequencing-based mapping of populations now being possible, new approaches for trait mapping using two contrasting bulks for the given traits have also been used.

Once molecular markers linked with traits are identified, they can be used for **marker-assisted back crossing (MABC)** or marker-assisted selection (MAS) programs [7]. ADSTs can be helpful in selecting superior lines based on foreground and background selection for the next crossing. The other two approaches of GAB also require ADSTs, specifically **marker-assisted recurrent selection (MARS)**, which enables the accumulation of superior alleles from different genetic backgrounds to one background, and **genomic selection (GS)**, which enables enhancing genetic gain in crop breeding. Furthermore, data generated during the course of one GAB program often need to be shared with different partners to better enable future GAB programs in other institutes and countries. Therefore, ADSTs are required for the management, retrieval, and sharing of data.

In view of all of the above, it is evident that appropriate ADSTs and their integrated use at the right time in different steps of GAB is critical for the next generation of genomics and integrated breeding (see Outstanding Questions). This review discusses the need, availability, and future requirements of ADSTs for enhancing the precision and modernizing of crop breeding (Table S1 in the supplemental information online and Figure 1).

Genetic Diversity and Population Genetic Analysis

Genetic diversity estimates help to structure germplasm defining, for example, heterotic pools, and provide useful information to select contrasting parental lines for new breeding populations. Analysis of molecular marker-based estimates of genetic diversity depends on a number of criteria, such as type (dominant or codominant) of markers, number of markers and genotypes, missing data, and proportion of heterozygosity. Similarly, population genetic analysis provides estimates of the allele frequencies that are helpful to breeders because alleles are the raw material for selection in breeding programs [8]. High-density genotyping has revolutionized the identification of favorable alleles in populations, minimizing the risk of recombination between markers and target genes.

To analyze genetic diversity, the numerical taxonomy and multivariate analysis system (NTSYSpc) is one of the most widely used software tools (see Table S1 for a description of all of these tools) [9]. Molecular evolutionary genetic analysis (MEGA) is another widely used program for estimating evolutionary distances and phylogenetic trees from DNA or protein sequence data [10]. Although several other programs/tools are available, DARwin (http://darwin.cirad.fr), a free and easy to use program for diversity and multivariate analysis of datasets that also provides publication-ready figures, has emerged as a popular tool in recent years for genetic diversity analysis. Due to the capacity to generate millions of SNPs in germplasm collections, construction of phylogenetic trees is an increasingly computationally challenging task. In this context, new pipelines such as SNPhylo [11] are being developed. SNPhylo works by selecting one informative SNP from each **linkage disequilibrium (LD)** block, thereby greatly decreasing the running time without losing much information.

For population genetics analysis, Arlequin is a highly used software package for molecular variance (AMOVA) analysis of datasets that includes several statistics like diversity, genetic

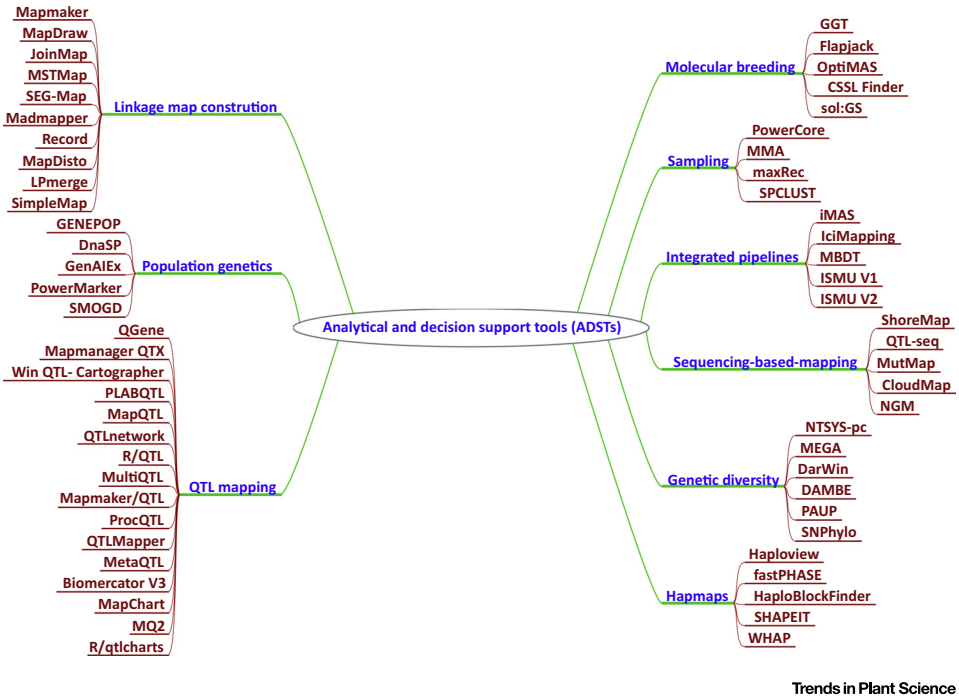

Figure 1. The Most Popular Analytical and Decision Support Tools (ADSTs) Used for Specific Purposes. This figure presents various analytical and decision support tools for genomics-assisted breeding components including linkage map construction, population genetic analysis, quantitative trait locus (QTL) mapping, molecular breeding, sampling, integrated pipelines, sequencing-based mapping, genetic diversity, and hapmaps. Different ADSTs can be selected based on their suitability for the experiment along with the strength of the tools. The details of individual ADSTs are presented in Table S1 in the supplemental information online.

distance, equilibrium analysis, and neutrality tests [12]. DNA Sequence Polymorphism (DnaSP) utilizes DNA sequence data and estimates several measures of DNA sequence variation within and between populations including LD, recombination, gene flow, and gene conversion parameters and can perform several tests of neutrality [13]. GenAlEx, which is based on Microsoft Excel, offers a wide range of population genetic analysis options for the full spectrum of genetic markers with rich graphical outputs for data exploration and publication [14]. Several other software tools have become available in recent years for various applications. For example, Power Marker is useful for simple sequence repeat (SSR) or SNP marker datasets for population genetic analysis and has a user-friendly graphical interface [15].

Based on sequence/marker diversity analysis on a large scale, a core set of germplasm, also called a reference set [16], can be developed. Reference sets seem to be better than core collections (comprising ~10% of the entire collection [4]) and mini-core collections (comprising ~10% of the total core collection or 1% of the total collection [5] for undertaking GWAS, as discussed below), as they have lower structural components than the full germplasm sets. Furthermore, the concept of selective phenotyping is also increasingly popular for selecting the subsets of mapping populations. This type of mapping is often done using recombination breakpoints to eliminate the need to extensively phenotype large numbers of individuals [17]. This is important in the case of populations like **multiparent advanced generation intercross (MAGIC)** [a population developed by crossing multiple founder lines (four or eight) to improve the precision and resolution of QTL mapping] where large numbers of lines are available and genotyping can be done in a high-throughput manner, but phenotyping of such large number of lines is challenging.

Glossary

Analytical and decision support tools (ADSTs): refer to a wide range of computer-based tools (simulation models, algorithms, techniques, and/or methods) developed for the analysis of different datasets and selection of promising genotypes in GAB programs for the development of new breeding lines.

Bulked segregant analysis (BSA): an approach for the identification of molecular markers associated with the trait of interest through the genetic analysis of two different pools based on phenotypic extremes from the segregating population.

Chromosome segment substitution lines (CSSLs): these are powerful QTL mapping populations that are used to identify favorable alleles from unadapted germplasm. These are a series of near-isogenic lines in which each CSSL carries specific chromosome segments in the genetic background of the recipient parent.

Composite interval mapping (CIM): a combination of interval mapping with multiple regressions that separate individual QTL effects. It prevents genetic variation in other regions of the genome, which effects QTL detection.

Consensus map: developed through combining multiple genetic maps available for the same species to obtain a higher density of markers for greater genome coverage than any individual genetic map.

Core collection: a limited set of accessions (10% of the entire collections) that represents the maximum diversity of the entire set with a minimum of repetitiveness. A core collection is suitable germplasm set for allele mining and LD analysis.

Genetic map: the arrangement or ordering of genes/loci on the basis of recombination frequency on a chromosome by defining linkage groups.

Genome-wide association study (GWAS): a population-based statistical association analysis for the identification of marker trait associations based on LD through genotyping and phenotyping of diverse individuals.

Genomics-assisted breeding (GAB): a method of breeding in which the selection of genotypes depends on genome information including molecular markers. More

For a selection of genotypes for core or mini-core formation, PowerCore is a widely used software package. It was developed based on advanced Maximization (M) strategy with a heuristic search for establishing core sets [18]. The M strategy has been used to select specific combinations of accessions that include complete coverage and is useful for selecting entries with the most diverse alleles and eliminating redundancy. It has been suggested that before considering molecular markers datasets for the construction of core sets, the data resolution (DR) needs to be calculated using a jackknife approach to the selection of suitable marker sets [19]. However, for selection of lines (with maximal dissimilarity) from the mapping population for undertaking selective phenotyping, three main methods are available. The minimum moment aberration (MMA) method minimizes the average of all pairwise similarities between the individuals of the population. It can be utilized in selecting F_2 recombinants without any missing datasets [20]. maxRec is another statistical tool for selecting lines on the basis of higher numbers of recombination events during the course of the recombination generations [21]. This statistical package is suitable for backcross, double haploid, and recombinant inbred line (RIL) populations. SPCLUST is another program that has been developed for selecting lines from BC, F_2 intercross, and complex crosses like four-way MAGIC [19]. The power of QTL detection using selected subsets using SPCLUST was similar to the power that could be achieved by using the entire dataset for analysis for QTL experiments.

Construction of Genetic Maps

Genetic maps serve as the foundation for various genetic applications, such as ordering of genes/markers, QTL mapping, association mapping, and map-based cloning [22]. Genetic maps are useful for anchoring scaffolds to linkage groups as well as assembling (and sometimes correcting) smaller contigs into large contigs [1]. However, construction of high-quality genetic maps depends on the following four parameters: the type of the population (e.g., biparental populations like F_2, $F_{2:3}$, BC, RILs, NILs, DH, multiparental mapping populations); the size of the population (100–500 lines); the number of markers (50–100 000); and the nature of the markers (SSR, DArT, SNP). Managing all of these parameters requires skills like working on a LINUX platform as well as high-performance computing programs [23].

Construction of linkage maps for small-scale experiments with fewer markers (<500) and smaller population sizes (<200) can still be undertaken with the first-generation and most widely used mapping software tool MAPMAKER [24]. MapDraw is a simple Microsoft Excel-based free software tool that can create attractive linkage maps as well as undertaking various kind of analysis [25]. JoinMap (https://www.kyazma.nl/index.php/JoinMap/) is a Windows-based software tool that can handle up to 50 000 markers and its key capability is to integrate data from multiple populations. This software generates high-quality publication-ready images. Recombination Counting and Ordering (Record) is a statistical tool that can be utilized for ordering marker loci on genetic maps [26]. Recently, an ultrafast pipeline, namely SimpleMap (http://simplemap-aj.sourceforge.net/), was streamlined for the construction of high-density linkage maps. This pipeline can develop linkage maps with ~1000 loci in <10 min, compared with >8–10 h using other programs.

Currently, genotype data is becoming available for 50 000 to 100 000 marker loci. For such marker densities, MSTMap has been developed and works on a minimum spanning tree (MST)-based method (http://www.mstmap.org/). The MST algorithm uses well-established graph theory and provides an efficient solution to the generation of genetic maps using large numbers of markers and individuals. MSTMap outperforms other mapping programs when the input data are noisy or incomplete. To manage large-scale re-sequencing data on the population, the sequencing enabled genotyping based map (SEG-Map) has also been developed for the construction of linkage maps [27]. This software allows the mapping of short reads generated for progeny into pseudomolecules of the parents of the mapping population, which in turn

precisely, GAB is the application of various genetic and genomics tools to develop new breeding lines.

Genomic-estimated breeding values (GEBVs): estimated breeding values generated through genotyping of populations using statistical model (s) and used to select superior individuals in a segregating population.

Genomic selection (GS): is a new method of molecular breeding in which selection of lines is based on GEBVs calculated based on genome-wide markers. GEBVs can be estimated through genotyping and phenotyping of a training population.

High-throughput genotyping (HTPG): a powerful and efficient method for rapid analysis of DNA sequence variations among large number of samples using the most advanced techniques, thereby generating a huge set of datasets that can be analyzed to understand nucleotide variations.

Linkage disequilibrium (LD): nonrandom association between two markers, genes or, QTLs on the same chromosome in a population owing to their tendency to be coinherited. When variants of two genetic loci are in LD, the variant seen at one locus predicts the variant found at the other.

Linkage drag: the carry-forward of any unwanted genes/loci along with the trait of interest from a donor parent during a backcross breeding program that might reduce the agronomic character of the elite cultivar.

Marker-assisted back crossing (MABC): the breeding method for introgression of major effect loci (two to four) in an elite genetic background through marker-aided foreground selection (selection of plants with the desired alleles from the donor parent) and supplemented with background (selection of plants with higher recurrent parent genome) in a rapid and precise manner.

Marker-assisted recurrent selection (MARS): a marker-based breeding process used to identify and monitor key regions (up to 20 or more) from both of the superior parents for complex traits in consecutive breeding generations.

Mini-core collection: a limited set of accessions (about 10% of a core collection or 1% of the entire collection) without losing much genetic diversity. The smaller size of

enables detection of SNPs. These SNPs can be used to identify recombination breakpoints and for bin map construction. The output data of SEG-Map can be directly used for QTL mapping studies.

For a given species, several genetic maps have sometimes been developed using various mapping populations. As a result, no single genetic map has a marker order for all markers available in that crop, and sometimes maps from different populations are different. Consensus genetic maps based on multiple biparental mapping populations are, therefore, an important resource for providing order for large numbers of marker loci for a given species. These maps are useful for analyzing LD as well as for association analysis and fine mapping of QTLs. Based on the availability of the common markers mapped from different mapping populations, **consensus maps** have been generated in many crops using the JoinMap program [28–31]. Recently, LPmerge, a new R-based package, has also been developed to construct consensus maps, with a major focus on marker orders to remove and resolve the conflicts in consensus maps [32]. Programs like JoinMap, MSTMap, and SEG-Map are increasingly common for the construction of high-density maps. Similarly, JoinMap or LPmerge will be useful for the development of consensus maps from different mapping populations. At present, medium numbers of markers (200–500) are being used for linkage map development and trait mapping. However, in the future, with the advent of sequencing-based trait mapping, current programs/methods for the development of high-density linkage maps may become obsolete. With the rapid development of sequencing technologies and the possibility to sequence hundreds/thousands of accessions at species or even genus level, a pan-genome for the species/genus can be developed. Such pan-genomes have already been developed in some crops like maize [33], rice [34], and soybean [35]. The hapmap information coming from these pan-genomes should serve as the foundation for the construction of 'universal maps' for given species/genera.

Linkage-Mapping Based QTL Analysis

QTL mapping, in general, uses one of following approaches: **single marker analysis (SMA)**, **simple interval mapping (SIM)**, or **composite interval mapping (CIM)**. However, it can be further extended in terms of estimating epistatic and environmental interactions [36,37]. Most QTL mapping tools have been developed for biparental mapping populations (Table S1). However, in recent years some sophisticated tools have been developed for multiparent mapping populations, like MAGIC and **nested association mapping (NAM)** populations.

Although a range of QTL analysis programs are available, QGene [38], MapManager QTX (http://iubio.bio.indiana.edu/soft/molbio/mac/map-manager-readme.html), and MapMaker/QTL (http://www.broadinstitute.org/ftp/distribution/software/mapmaker3/) are the appropriate software tools for SMA. For CIM, WinQTL Cartographer (http://statgen.ncsu.edu/qtlcart/WQTLCart.htm), MapQTL (https://www.kyazma.nl/index.php/mc.MapQTL), and PLABQTL [39] have been shown to be appropriate software [40]. Inclusive CIM (ICIM) [41] and QTLNetwork [42] are other commonly used programs for QTL mapping.

To analyze marker–trait associations (MTAs) and to finely map genetic regions in multiparent mapping populations of outbred animal stocks, the specialized software package HAPPY was developed [43]. However, in the case of plant species, R-based packages such as R/qtl, R/ricalc, R/mpMap, and R/mpwgaim have been used to map genomic regions [44–46]. For analyzing multiparent mapping populations like NAM populations, an integrated software tool called IciMapping (see detailed description later) has been developed to identify the genomic regions responsible for the trait of interest [41]. As many QTL analysis studies are based on different populations with phenotyping data from different environments, many researchers have started to undertake meta-QTL analysis to understand the genetic determination of complex traits. This approach is also useful for the identification of robust QTLs, which can be subjected

mini-core collections makes them suitable for a wide range of applications including trait mapping and breeding.

Multiparent advanced generation intercross (MAGIC): a type of population developed by crossing multiple founder lines (four or eight) to improve the precision and resolution of QTL mapping. Multiparent crossing creates a mosaic of the founders through reshuffling of the genome of each parental line, enabling fine mapping of the QTLs at a higher genetic resolution.

Nested association mapping (NAM): combines the advantages of linkage and association mapping and eliminates the disadvantages of both methods. NAM takes into consideration both recent and historical recombination events to define the genomic region responsible for the trait of interest, with high mapping resolution.

Quantitative trait locus (QTL): a genomic region possessing several minor- and/or large-effect genes on a chromosome and responsible for complex quantitative traits.

Reference set: usually developed based on the molecular characterization of composite collection (which may include core and mini-core collections). These sets are ideal for genetic diversity analysis, population structure, and association mapping.

Simple interval mapping (SIM): testing for the presence of QTLs at many positions, between each pair of adjacent markers. The SIM method calculates a LOD score, on the basis of which the probability of the presence of a QTL at that position can be indicated.

Single marker analysis (SMA): a part of QTL analysis where associations between molecular markers and traits of interest can be detected using a single marker at a time by calculating the recombination frequencies of linked genes.

Whole-genome re-sequencing: sequencing of the genomes of individual lines for the species for which the reference genome is available. Provides a wide range of variants, mutations, structural variation, copy number variation, and rearrangements between and among individuals.

for fine mapping and ultimately useful for the identification of candidate genes. To perform meta-analysis of QTLs, MetaQTL [47] and BioMercator [48] are promising software packages. Additionally, to develop linkage maps and to project QTLs, several other packages have become available recently, including MapChart [49], MQ^2 [50], and R/qtlcharts [51].

GWAS

In the case of GWAS, understanding population structure and the level and distribution of LD in the populations is a prerequisite for using the appropriate approach of association mapping. In this context, STRUCTURE [52] is the most extensively used software to detect population genetic structure. STRUCTURE generates clusters based on both transient Hardy–Weinberg disequilibrium and LD caused by admixture between populations [53,54]. EIGENSOFT is another widely used statistical package for the detection and correction of population stratification in GWAS using principal component analysis [55]. Similarly, Bayesian analysis of population structure (BAPS) is another program for Bayesian inference of the genetic structure, especially for analyzing large-scale population genetics data in a population [56]. Furthermore, for analysis of re-sequencing data in terms of LD and haplotype block analysis, haplotype population frequency estimation, single SNP and haplotype association tests, and permutation tests for association significance, SNP analyzer 2.0 has been developed [57]. A detailed list of other available software for LD analysis can be found at http://www.genes.org.uk/software/LD-software.shtml.

For performing association analysis, Trait Analysis by aSSociation, Evolution, and Linkage (TASSEL) is the most commonly used and highly cited software in GWAS in plants [58]. This software provides several new and powerful statistical approaches for association mapping such as the General Linear Model (GLM) and Mixed Linear Model (MLM) [59]. GenABEL (http://www.genabel.org/manuals/GenABEL) is a genome-wide SNP-association analysis program based on R. PLINK is another highly cited open-source software for whole-genome association analysis. This program is designed to perform a range of basic and large-scale analyses [60]. PLINK focuses on the analysis of genotype/phenotype data to perform the association analysis.

Most association mapping analyses have been conducted based on GLM or MLM, which does not seem sufficient for the identification of robust MTAs. Therefore, in the near future, models such as multilocus mixed models (MLMMs) and multitrait mixed models (MTMMs) need to become more common. To confirm the association of SNPs identified from the GWAS with the target traits, nonsynonymous SNP (nsSNP)-based association mapping is one of the most promising approaches [61]. Additionally, with the increasing use of small Indels for MTAs, the ADSTs available at present need to be modified in such a way that they can accommodate SNPs as well as small Indels for performing trait association analysis. Recently, NGS-based trait mapping approaches and analysis of re-sequencing data were found promising for the identification of target genomic regions. In this context, large numbers of scripts/software were developed that could be deployed in NGS-based studies, including haplotype-based GWAS (Box 1). For better utilization of GWAS results, the identified MTAs in various crops should be made available as open-access databases for the selection and deployment of the most robust alleles in crop improvement programs.

Molecular Breeding

Significant progress has been achieved in the area of molecular breeding in developing improved plant varieties [62,63]. Among various GAB approaches, MAS/MABC has been used extensively in public breeding programs. MAS/MABC, in general, do not use any sophisticated tools to select plants for advancement or backcrossing. However, open-access visualization tools such as Graphical Genotypes (GGT) [64], Flapjack [65], and the molecular breeding design tool

Box 1. Genomic Tools for Sequencing-Based Mapping and Re-sequencing Analysis

NGS-based mapping approaches using **bulked segregant analysis (BSA)** have been used for mapping target genomic regions without the construction of linkage maps. However, these approaches require specialized skills and tools. Some approaches have been developed for facilitating trait mapping using NGS approaches. For instance, the ShoreMap approach (simultaneous mapping and mutant identification by deep sequencing) was developed to map the target genes in mutant lines [76]. This software package, available at http://1001genomes.org/software/shoremap.html, is open source and is continuously updated. Similarly, the next-generation mapping (NGM) (http://bar.utoronto.ca/ngm/) pipeline was proposed and developed for trait mapping [77].

The MutMap [78] and QTL-seq [79] approaches facilitate the mapping of targeted genomic regions from EMS-derived mutants and from any desirable genotype, respectively. To perform either of these two analyses, specific bioinformatics pipelines are available at http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap. CloudMap (http://usegalaxy.org/cloudmap) is another open-source web-based analytical bioinformatics pipeline for the identification of candidate genes directly from EMS-derived mutants without the development of a mapping population [80].

Re-sequencing of numbers of lines from different crop species opens new avenues and is useful for understanding the evolution of and genetic relationships among individuals. Therefore, for analyzing re-sequencing datasets in terms of haplotypes, construction of hapmaps, haplotype population frequency estimation, single SNP and haplotype association tests, and permutation tests for association significance, Haploview is a promising tool. Haploview can analyze thousands of SNPs (tens of thousands in command-line mode) in thousands of individuals [81]. SHAPEIT [82], fastPHASE [83], WHAP [84], and HaploBlockFinder [85] are some other important analytical tools/pipelines for the development of hapmaps/haplotypes and performing GWAS.

(MBDT) (https://www.integratedbreeding.net/179/training/bms-user-manual/marker-assisted-backcross-breeding-tool) have become available in recent years for the selection of plants with maximum recurrent parent genome recovery at the global level to eliminate the precise **linkage drag** on carrier chromosomes. Another data visualization and selection tool called CSSL Finder is useful for developing **chromosome segment substitution lines (CSSLs)** [66]. This is a useful tool to search a population of advanced backcross lines for a set of lines with the optimized representation of the donor parent genome in the recurrent parent background. This software, in conjunction with a graphical genotype, also displays the phenotypic values of the individual lines. Therefore, this program is useful for the identification of elite/novel CSSLs responsible for a trait of interest.

For MARS, OptiMAS has been developed by the French Agricultural Research Centre for International Development (CIRAD) as a part of the Integrated Breeding Platform (IBP). This software helps in selecting plants possessing superior alleles from elite parents in several cycles of recombination [67]. GS is a new molecular breeding approach that integrates marker data and phenotypic data from a training population to generate a prediction model for predicting **genomic-estimated breeding values (GEBVs)** for all segregating individuals of a breeding population. Calculation of GEBV requires specific statistical models that treat markers as random effects. The most commonly used GS prediction models are the Random Regression Best Linear Unbiased Predictor (RR-BLUP) [68], BayesA [68], BayesB [68], BayesCπ [68], Bayesian Ridge Regression (RR) [69], Bayesian LASSO [70,71], and Random Forest Regression (RFR) [72]. However, no single statistical model has emerged as being clearly better than the others for all applications. For some applications, it is possible to select the most suitable model after testing several alternative models. In this context, solGS, a web-based tool for GS based on the RR-BLUP model, has been developed [73]. This software is an easy-to-use analysis platform for performing GS in plant breeding. Similarly, ISMU 2.0 is being developed by ICRISAT, with the close collaboration of several leading institutions. ISMU 2.0, which is an improved version of ISMU 1.0 [74], has several data processing capabilities including several models of GS. This pipeline includes most of the GS models, including RR-BLUP, Kinship Gauss, RR, Bayesian LASSO, BayesA, BayesB, BayesCπ, and RF and works on Windows, CentOS, and Ubuntu platforms [75]. Thus, ISMU 2.0 will be useful for the breeding community to analyze large-scale datasets for GS experiments for enhancing genetic gains (Box 2).

Box 2. Integrated Pipelines for GAB

Most of the tools presented in this review are standalone applications that use different, and not necessarily compatible, formats, especially for their input and ouput files. Their different technical charateristics and specifications represent a major constraint on their use in routine breeding activities. To overcome this bottleneck, a few platforms are emerging that offer continuous analytical and decision-making pipelines, integrating ADSTs in a seamless fashon whereby the ouput of one tool in the pipeline is readily accepted as input by the next tool in the chain.

To identify the MTA using biparental mapping population, integrated MAS (iMAS) was developed by ICRISAT. It is an open-source integrated molecular breeding analysis platform to facilitate trait mapping based on freely available and powerful software tools (http://www.icrisat.org/bt-software-imas.htm). This software suite comprises six different modules including data validation, phenotypic evaluation, linkage map construction, QTL analysis, QTL projection, and marker-assisted breeding.

The Integrated Breeding Platform (IBP) (http://www.integratedbreeding.net), developed recently by the CGIAR's Generation Challenge Programme and partners, is a web-based one-stop shop for information, analytical tools, and related services to design and carry out integrated breeding projects. The IBP aims to provide, on a single portal, access to the crop information, tools, and services that a breeder needs to conduct modern genomics- and/or informatics-based breeding activities. The core 'product' of the IBP is the Breeding Management System (BMS), an integrated application of various data management, statistical analysis, and decision support tools to support the various stages of the crop breeding process toward the release of improved germplasm [86]. The BMS provides useful tools for analyzing phenotypic and genotypic datasets and managing day-to-day activities through all phases of breeding programs. This is open-source and one-stop shopping for all of the tools required for GAB programs. One such tool for MAS experiments is MBDT (https://www.integratedbreeding.net/179/training/bms-user-manual/marker-assisted-backcross-breeding-tool). MBDT comprises of six modules including data validation, phenotyping, linkage map building, QTL analysis, genome display, and MABC sample size.

With an objective to identify and use of SNPs in breeding programs, ICRISAT recently developed a pipeline called Integrated SNP Mining and Utilization (ISMU 1.0) for preprocessing of the raw NGS data, SNP detection between any combinations of genotype, visualization of the alignment and SNPs results, and the development of KasPar and Golden Gate assays for a range of experiments [74]. This pipeline has been extended to ISMU 2.0 [75]. The updated pipeline comprises several genomic selection modules for estimating GEBVs and the selection of superior lines.

Strategic Outlook on the Future Prospects

While significant advances have been made in the areas of genomics and GAB, further efforts need to be made to develop ADSTs and crop information systems. In the area of data management for crop breeding (storage, curation, analysis, and publication) one size clearly does not fit all, so there is an increasing need to better integrate software tools and develop interoperable application program interfaces (APIs) to facilitate access to diverse tools and databases across different pipelines. In addition to these analytical and bioinformatics needs, we identify here four other areas that could be addressed to improve the efficiency of GAB: (i) further reduction in genotyping costs per line so that genome-wide marker profile data can be generated on large populations in routine breeding activities for enhancing/fixing favorable alleles; (ii) reduction in field-relevant phenotyping costs and implementation of new high-throughput screening methods such as aerial infrared screening; (iii) adoption of best, or at least good, data management practices, starting with adequate resource allocation and implementation of a data management policy at institute level through the joint efforts of management and the donor community; and (iv) a sustainable adoption of ADSTs and associated programs/tools that goes beyond just technological development to include training and suitable support services for breeders.

Deployment of ADSTs and crop information systems must be considered carefully. Some of these issues need to be adressed through an integrated approach and one should not underestimate the difficulty related to technology transfer in the public sector. Local support, such as that provided by the IBP through regional hubs, can be an attractive option to enhance the use of modern breeding tools and services. This would be mainly through capacity building, technical support, and crop-specific expertise.

We are hopeful that the development and deployment of the right ADSTs at the right time, in keeping with the needs, resources, and technical readiness of breeding programs, will usher

crop improvement programmes into a modern, knowledge-based crop improvement era, leading to sustainable crop production and global food security.

Acknowledgments

R.K.V. thanks the CGIAR Generation Challenge Programme (GCP), the US Agency for International Development (USAID), and the Department of Biotechnology, Government of India for sponsoring research at ICRISAT on the topics mentioned in the review. This study has been undertaken as part of the CGIAR Research Program on Grain Legumes. ICRISAT is a member of the CGIAR Consortium.

Supplemental Information

Supplemental information associated with this article can be found, in the online version, at doi:10.1016/j.tplants.2015.10.018.

Outstanding Questions

What are the major ADSTs with specific features that are available for plant breeding?

Will the construction of genetic maps be obsolete and/or outdated in the context of generating millions of data points on segregating populations?

Can pan-genome information for a given crop be useful for developing high-density linkage maps that can serve as 'universal maps'?

Can open-source and one-stop integrated platforms facilitate GAB programs in developing countries?

References

1. Varshney, R.K. *et al.* (2013) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* 30, 83–99
2. Varshney, R.K. *et al.* (2014) Marker assisted backcrossing to introgress resistance to *Fusarium* wilt (FW) race 1 and *Ascochyta* blight (AB) in C 214, an elite cultivar of chickpea. *Plant Genome* 7, 11
3. Xu, Y. (2010) *Molecular Plant Breeding,* CAB International
4. Brown, A.H.D. (1989) Core collections: a practical approach to genetic resources management. *Genome* 31, 818–824
5. Upadhyaya, H.D. and Oritz, R. (2001) A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. *Theor. Appl. Genet.* 102, 1292–1298
6. Upadhyaya, H.D. *et al.* (2008) Genetic structure, diversity, and allelic richness in composite collection and reference set in chickpea (*Cicer arietinum* L.). *BMC Plant Biol.* 8, 106
7. Varshney, R.K. *et al.* (2012) Can genomics boost productivity of orphan crops? *Nat. Biotechnol.* 30, 1172–1176
8. Labbate, J.A. (2000) Software for population genetic analyses of molecular marker data. *Crop Sci.* 40, 1521–1528
9. Rohlf, F.J. (1992) *NTSYS-pc (Numerical Taxonomy and Multivariate Analysis System). Version 1.70,* Exeter
10. Tamura, K. *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739
11. Lee, T.H. *et al.* (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* 15, 162
12. Excoffier, L. *et al.* (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* 1, 47–50
13. Librado, P. and Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452
14. Peakall, R. and Smouse, P.E. (2012) GenAlex 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics* 28, 2537–2539
15. Liu, K. and Muse, S.V. (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21, 2128–2129
16. Glaszmann, J.C. *et al.* (2010) Accessing genetic diversity for crop improvement. *Curr. Opin. Plant Biol.* 13, 167–173
17. Huang, B.E. *et al.* (2012) Selecting subsets of genotyped experimental populations for phenotyping to maximize genetic diversity. *Theor. Appl. Genet.* 126, 379–388
18. Kim, K.W. *et al.* (2007) PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23, 515–526
19. van Hintum, T.J. (2007) Data resolution: a jackknife procedure for determining the consistency of molecular marker datasets. *Theor. Appl. Genet.* 115, 343–349
20. Jin, C. *et al.* (2004) Selective phenotyping for increased efficiency in genetic mapping studies. *Genetics* 168, 2285–2293
21. Jannink, J.L. (2005) Selective phenotyping to accurately map quantitative trait loci. *Crop Sci.* 45, 901–908
22. Wu, Y. *et al.* (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.* 4, e1000212
23. Cheema, J. and Dicks, J. (2009) Computational approaches and software tools for genetic linkage map estimation in plants. *Brief. Bioinform.* 10, 595–608
24. Lander, E.S. *et al.* (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1, 174–181
25. Liu, R. and Meng, J. (2003) [MapDraw: a Microsoft Excel macro for drawing genetic linkage maps based on given genetic linkage data]. *Yi Chuan* 25, 317–321 (in Chinese)
26. Van Os, H. *et al.* (2005) RECORD: a novel method for ordering loci on a genetic linkage map. *Theor. Appl. Genet.* 112, 30–40
27. Zhao *et al.* (2010) SEG-Map: a novel software for genotype calling and genetic map construction from next-generation sequencing. *Rice* 3, 98–102
28. Somers, D.J. *et al.* (2004) A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 109, 1105–1114
29. Varshney, R.K. *et al.* (2007) A high density barley microsatellite consensus map with 775 SSR loci. *Theor. Appl. Genet.* 114, 1091–1103
30. Gautami, B. *et al.* (2012) An international reference consensus genetic map with 897 marker loci based on 11 mapping populations for tetraploid groundnut (*Arachis hypogaea* L.). *PLoS ONE* 7, e41213
31. Shirasawa, K. *et al.* (2013) Integrated consensus map of cultivated peanut and wild relatives reveals structures of the A and B genomes of *Arachis* and divergence of the legume genomes. *DNA Res.* 20, 173–180
32. Endelman, J.B. and Plomion, C. (2014) LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics* 30, 1623–1624
33. Hirsch, C.N. *et al.* (2014) Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* 26, 121–135
34. Schatz, M.C. *et al.* (2014) Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. *Genome Biol.* 15, 506
35. Li, Y.H. *et al.* (2014) *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* 32, 1045–1052
36. Lander, E.S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121, 185–199
37. Manly, K.F. *et al.* (2001) Map Manager QTX, cross-platform software for genetic mapping. *Mamm. Genome* 12, 930–932
38. Joehanes, R. and Nelson, J.C. (2008) QGene 4.0, an extensible Java QTL-analysis platform. *Bioinformatics* 24, 2788–2789
39. Utz, H.F. and Melchinger, A.E. (1996) PLABQTL: a program for composite interval mapping of QTL. *J. Agric. Genomics* 2, 1–4

40. Varshney, R.K. *et al.* (2009) Molecular plant breeding: methodology and achievements. *Methods Mol. Biol.* 513, 283–304
41. Li, H. *et al.* (2008) Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. *Theor. Appl. Genet.* 116, 243–260
42. Yang, J. *et al.* (2008) QTLNetwork: mapping and visualizing genetic architecture of complex traits in experimental populations. *Bioinformatics* 24, 721–723
43. Mott, R. *et al.* (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci. U.S.A.* 97, 12649–12654
44. Broman, K.W. *et al.* (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890
45. Broman, K.W. (2005) The genomes of recombinant inbred lines. *Genetics* 169, 1133–1146
46. Huang, B.E. and George, A.W. (2011) R/mpMap: a computational platform for the genetic analysis of multi-parent recombinant inbred lines. *Bioinformatics* 27, 727–729
47. Veyrieras, J.B. *et al.* (2007) MetaQTL: a package of new computational methods for the meta-analysis of QTL mapping experiments. *BMC Bioinformatics* 8, 49
48. Sosnowski, O. *et al.* (2012) BioMercator V3: an upgrade of genetic map compilation and quantitative trait loci meta-analysis algorithms. *Bioinformatics* 28, 2082–2083
49. Voorrips, R.E. (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93, 77–78
50. Chibon, P.Y. *et al.* (2013) MQ2: visualizing multi-trait mapped QTL results. *Mol. Breed.* 32, 981–985
51. Broman, K.W. (2014) R/qtlcharts: interactive graphics for quantitative trait locus mapping. *Genetics* 199, 359–361
52. Pritchard, J.K. *et al.* (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959
53. Porras-Hurtado, L. *et al.* (2013) An overview of STRUCTURE: applications, parameter settings and supporting software. *Front. Genet.* 4, 1–13
54. Kalinowski, S.T. and Powell, J.H. (2015) A parameter to quantify the degree of genetic mixing among individuals in hybrid populations. *Heredity* 114, 249–254
55. Price, A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909
56. Corander, J. and Marttinen, P. (2006) Bayesian identification of admixture events using multilocus molecular markers. *Mol. Ecol.* 15, 2833–2843
57. Yoo, J. *et al.* (2008) SNPAnalyzer 2.0: a web-based integrated workbench for linkage disequilibrium analysis and association analysis. *BMC Bioinformatics* 9, 290
58. Bradbury, P.J. *et al.* (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635
59. Zhang, Z. *et al.* (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360
60. Purcell, S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575
61. Wellcome Trust Case Control Consortium *et al.* (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* 39, 1329–1337
62. Gupta, P.K. *et al.* (2010) Marker assisted selection as a component of conventional plant breeding. *Plant Breed. Rev.* 33, 145–217
63. Varshney, R.K. *et al.* (2015) Translational genomics in agriculture: some examples in grain legumes. *Crit. Rev. Plant Sci.* 34, 169–194
64. Van Berloo, R. (2008) Computer note: GGT 2.0: versatile software for visualization and analysis of genetic data. *J. Hered.* 99, 232–236
65. Milne, I. *et al.* (2010) Flapjack – graphical genotype visualization. *Bioinformatics* 26, 3133–3134
66. Lorieux, M. (2012) MapDisto: fast and efficient computation of genetic linkage maps. *Mol. Breed.* 30, 1231–1235
67. Valente, F. *et al.* (2013) OptiMAS: a decision support tool for marker-assisted assembly of diverse alleles. *J. Hered.* 104, 586–590
68. Meuwissen, T.H.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829
69. de los Campos, G. *et al.* (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385
70. Habier, D. *et al.* (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 23, 186
71. Legarra, A. *et al.* (2011) Improved LASSO for genomic selection. *Genet. Res. (Camb.)* 93, 77–87
72. Breiman, L. (2001) Random forests. *Mach. Learn.* 45, 5–32
73. Tecle, I.Y. *et al.* (2014) solGS: a web-based tool for genomic selection. *BMC Bioinformatics* 15, 398
74. Azam, S. *et al.* (2014) An integrated SNP mining and utilization (ISMU) pipeline for next generation sequencing data. *PLoS ONE* 9, e101754
75. Rathore, A. *et al.* (2015) ISMU 2.0: a multi-algorithm pipeline for genomic selection. In *Plant and Animal Genome XXII, January 11-15, San Diego, CA*. Scherago
76. Schneeberger, K. *et al.* (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* 6, 550–551
77. Austin, R.S. *et al.* (2011) Next-generation mapping of *Arabidopsis* genes. *Plant J.* 67, 715–725
78. Abe, A. *et al.* (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* 30, 174–178
79. Takagi, H. *et al.* (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J.* 74, 174–183
80. Minevich, G. *et al.* (2012) CloudMap: a cloud-based pipeline for analysis of mutant genome sequences. *Genetics* 192, 1249–1269
81. Barrett, J.C. *et al.* (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265
82. Delaneau, O. *et al.* (2013) Improved whole chromosome phasing for disease and population genetic studies. *Nat. Methods* 10, 5–6
83. Scheet, P. and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644
84. Purcell, S. *et al.* (2007) WHAP: haplotype-based association analysis. *Bioinformatics* 23, 255–256
85. Zhang, K. and Jin, L. (2003) HaploBlockFinder: haplotype block analyses. *Bioinformatics* 19, 1300–1301
86. Delannay, X. *et al.* (2012) Fostering molecular breeding in developing countries. *Mol. Breed.* 29, 857–873